RESEARCH REPORT SERIES
(Disclosure Avoidance #2016-04)

Evaluating a Remote Access System

Michael H. Freiman, Bryan Schar, Kyle Hasenstab and Amy Lauger

Center for Disclosure Avoidance Research
U.S. Census Bureau
Washington DC 20233

Report Issued: July 29, 2016

Evaluating a Remote Access System

Michael H. Freiman

U.S. Census Bureau, Washington DC, United States

Bryan Schar

U.S. Census Bureau, Washington DC, United States

Kyle Hasenstab

Centers for Disease Control and Prevention, Atlanta, GA, United States

Amy Lauger

U.S. Census Bureau, Washington DC, United States

Summary. The U.S. Census Bureau has the dual aims of releasing useful data while protecting respondent confidentiality. We researched a system for custom tabulation queries of confidential microdata from the American Community Survey (ACS). The data in the system would be protected primarily by suppressing sparse tables. Research showed that the proposed rules would require an unacceptable number of table suppressions, while still being vulnerable to attacks on individual records. We concluded that our system as planned could not be useful while protecting the data adequately. Thus, we are considering other options to fulfill our aims.

Keywords: Confidentiality; Data access; Disclosure; Query system; Remote access

1. Introduction

The U.S. Census Bureau's data on the people and economy of the United States are in high demand from researchers, policymakers, businesses, non-profit organizations, students and other data users. However, Title 13 of the U.S. Code, which allows the Census Bureau to collect the information, also requires protection against the identification of a person, household or business in the data. This paper describes the Microdata Analysis System (MAS), which we researched as a potential way to make available as much useful data as possible while still protecting confidentiality.

The Census Bureau provides many data products and ways to access our data. Among the most accessible are the standard tabulations and profiles available through American FactFinder (AFF). The website contains billions of estimates at various levels of geography and is free to access online. However, AFF does not allow users to create custom tabulations or to combine tables. If a user wants a table for a custom area of three counties, the user can manually combine the numbers from three individual tables. However,

the user cannot calculate accurate margins of error for these combined estimates. Instead, the user can make assumptions about relationships between the estimates to approximate the margin of error.

Users who need estimates not included in the standard tables can request a custom tabulation. For the American Community Survey (ACS), such requests cost at least $3,000 and take at least eight weeks to process. The Census Bureau's Disclosure Review Board must approve all requests before tabulation can occur (U.S. Census Bureau, 2015a).

The Census Bureau releases public use microdata files containing individual record-level data for several surveys and censuses. The confidentiality protections for the files include de-identification, limiting geographic detail, coarsening, noise infusion, top coding and sometimes subsampling. Public use microdata records are identified only for predefined geographic areas with population 100,000 or more, and sometimes the populations are much larger. Hence, the files can be versatile but may not fulfill everyone's needs, especially when the user is interested in small geographic areas or the tails of a distribution.

The Census Bureau operates over 20 secure physical Federal Statistical Research Data Centers (FSRDCs), where approved users may analyze data from the Census Bureau and other government agencies. FSRDCs are another valued resource for researchers. However, using a FSRDC is costly and time-consuming, requiring an approved research proposal and background investigation before the research is performed, travel to the FSRDC and review of the results by Census Bureau staff to ensure they are safe to release.

After Census 2000, the Census Bureau released an online tool called the Advanced Query System (AQS). The AQS allowed a user to enter a custom query of decennial data and receive a table from the system. The system was only briefly available, but data users much desired its features.

Unless it is well-protected, a query system may allow a malicious user to manipulate multiple queries to discover confidential information about a person, household or business in the data. The AQS protected the data through a set of rules that each table had to pass before its release. Tables failing the rules were suppressed.

A more recent initiative at the Census Bureau is the Center for Enterprise Dissemination Services and Consumer Innovation (CEDSCI), which works to "enable the public to make better decisions using data through a continuously adaptive, customer-centric, open and accessible dissemination environment" (Blash, 2015). The MAS was planned as one piece of this framework, building on the AQS's features. The MAS would allow users to make a custom table of Census Bureau microdata, beginning with American Community Survey (ACS) data, ideally at a lower geographic level than the public use microdata areas (PUMAs) available in the ACS public use microdata. Users could create their own composite geographies from the geographic regions provided. Further capabilities, such as customized regressions and maps, could be added later. The MAS's functionality would be part of the larger CEDSCI platform, which will also provide other capabilities.

Section 2 discusses the confidentiality protections included in the MAS. Section 3 describes testing these protections to determine whether they would eliminate disclosure risk or limit the system's usefulness. From the testing, we conclude that the MAS as envisioned cannot balance these two considerations.

Section 4 explores in more detail how risk and utility are related when parameters of the system are varied. Because of the results of Sections 3 and 4, Section 5 discusses current research on synthetic data, another approach to releasing and protecting data which we are now focusing on more. Section 6 concludes by briefly describing the next steps in our research.

2. Protecting the MAS against disclosure

Tables in a query system may be protected by pre-tabulation or post-tabulation methods. Under a post-tabulation approach, the source data are perturbed minimally, and protection occurs after the analysis is performed. The system may perturb output, through methods such as semi-controlled random rounding, or may suppress tables deemed too risky to release. Under a pre-tabulation approach, the source microdata are directly perturbed or synthesized, providing protection to analyses performed on those data. A pre-tabulation approach may also be used when an agency provides microdata to the user outside of a query system. The two approaches are not mutually exclusive, but a system may put more emphasis on one than the other. The research described here focuses primarily on post-tabulation table suppression, even though we will mention other disclosure methods that we used in combination with it.

The MAS was to be available to any Internet user, and thus would require more disclosure protection than many other remote access systems. These other systems include Real Time Remote Access at Statistics Canada (Simard, 2011), the Remote Access Data Laboratory (RADL) at the Australian Bureau of Statistics (ABS) (2014a), LISSY from the LIS Cross-National Data Center in Luxembourg (2011) and ANDRE at the U.S. National Center for Health Statistics (NCHS) (Meyer, 2014). These systems often provide more versatility than the MAS would have or give output more frequently. This seeming laxity requires more restrictions in other areas. Many of these systems require a pre-approval process and vetting of results before release, making them akin to a virtual research data center. Systems providing custom tables without an approval process often give data at a higher level than we hoped to provide in the MAS, also requiring less protection of the data. Such table servers include ABS's DataAnalyser (Australian Bureau of Statistics, 2014b) and TableBuilder (Australian Bureau of Statistics, 2015), the former of which is in a limited-use beta version; the U.S. National Center for Education Statistics (NCES) DataLab and Data Analysis System (DAS) (National Center for Education Statistics, n.d. and n.d.); the NCHS Online Analytic Realtime System (OARS) under development (Meyer, 2015); and the U.S. Substance Abuse and the Mental Health Services Administration (SAMHSA) Restricted-use Data Analysis System (R-DAS) (Vsevolozhskaya and Anthony, 2014).

2.1 A differencing attack

Disclosure rules are necessary because a malicious user could manipulate tables based on our internal microdata file to identify a respondent. In a differencing attack, the user makes two tables that differ slightly, perhaps only by the inclusion of one unit in one of the tables. Subtracting one table from the other results in a constructed table that the system would not provide directly. Suppose the user has found a table of sex by age, where age has three categories. This table might look like Table 1. Conspicuously, only one man in the dataset is age 65 or older. The intruder may wish to find out more information about

this person, such as whether he is a veteran. Equivalently, the intruder would like to produce Table 2. The system would suppress this table, so the intruder takes an indirect approach, by making the Tables 3 and 4. Subtracting Table 4 from Table 3 results in Table 2 and shows that the target is a veteran.

Table 1: Sex by age

	Age 0-17	Age 18-64	Age 65+
Male	31	41	1
Female	59	26	53

Table 2: Veteran status for men age 65 and older

In armed services	Veteran	Not a veteran
0	1	0

Table 3: Veteran status for males

In armed services	Veteran	Not a veteran
6	9	58

Table 4: Veteran status for males under age 65

In armed services	Veteran	Not a veteran
6	8	58

Recovering enough variables could allow an intruder to reconstruct part of the target's microdata record. The intruder could then use the reconstructed data to find the target in the PUMS, potentially revealing the target's identity and characteristics. The system might deny the intruder the tables necessary to perform a differencing attack on veteran status, but in this case the intruder could try to use other variables to attack the record.

An attack could also result in the intruder's using the MAS to link Census data to an external dataset, which may include direct identifiers such as a person's name. Data that the Census Bureau had only made directly available in the aggregate or without direct identifiers could be linked to a record.

The opportunity to perform a differencing attack is not usually as obvious as it appears here, since the MAS would provide Table 1 only with weighted counts, and the unique observation ("unique") would rarely have a weight of 1. Hence, the intruder may have to find another way to confirm the cell includes a unique before proceeding. On the other hand, the intruder could potentially find the unique based on data available elsewhere, without constructing Table 1.

2.2 Post-tabulation disclosure avoidance through table suppression

Our approach to table suppression focused on three acceptance rules used in the AQS. Tables that did not pass the acceptance rules would be suppressed. The planned system would allow users to access only tables with survey weights. However, we applied the acceptance rules to an unweighted version of the table, since a major disclosure concern is table cells with few respondents in the sample. The rules were:

- The mean cell size must be at least a certain value *m*.
- The median cell size must be at least a certain value *n*.
- The proportion of cells with exactly one observation, among those cells with at least one observation, must not exceed a fixed value *p*.

The rules aimed to ensure that the system would suppress tables with small cell counts. The parameters used in the AQS and in testing the MAS are confidential.

In addition to suppressing some tables, the MAS would present variables as recodes with predefined categories, which the user could combine but not split. The system would check the acceptance rules based on the underlying non-combined categories. If allowed, splitting a recoded category could create a "sliver," a small subpopulation on which to perform a differencing attack. For example, if a user could create a recode for people with income $43,880 or more and another recode for people with income $43,881 or more, the user could create two tables and subtract to determine the attributes of the population with income exactly $43,880, which may be just one respondent.

We tested the system with the 2009-2013 ACS, the largest demographic survey dataset the Census Bureau produces. Although the ACS's design supports estimates at geographies as small as census block groups, our research was on geographies at least as large as census tracts, regions averaging 4,000 people. The source data already have undergone some forms of disclosure avoidance protection, including data swapping, in which pairs of households with similar characteristics have their geographic identifiers switched. Swapping affects only a small number of households and targets households deemed at high risk of disclosure (Lauger *et al.*, 2015).

2.3 A differencing attack based on geographic variables

Another form of differencing attack would occur if a user wanted a table on a geographic region that did not pass the acceptance rules. The user could request the same table on the desired region combined with an auxiliary region, perhaps dramatically larger then the initial one, then request the table on the auxiliary region alone. By subtracting the two results, a malicious user could derive virtually any table on the original region.

To prevent this possibility, we required that a table on a composite geography pass only if the same table would pass for each component geography taken individually. With this rule in place, tables on composite geographies create no additional disclosure risk but are also given less frequently than they would be otherwise.

3. Testing the acceptance rules' effectiveness

A query system such as the MAS must protect information about an individual record against disclosure while providing enough data to make the system useful.

3.1 The effectiveness of a differencing attack

The differencing attack described above will work only if the system produces Tables 3 and 4 and, if the intruder uses it, Table 1. We considered two possible variants of the rules when the universe is defined by both non-geographic and geographic conditions, as in Tables 2, 3 and 4.

Variant 1 (explicit universe definition): The system checks the rules against the unweighted equivalent of the table requested. In this example, the numbers checked for Table 3 are the unweighted totals corresponding to the numbers 6, 9 and 58 in Table 3.

Variant 2 (implied universe definition): Universe variables are treated as additional tabulation variables. In this example, a request for Table 3 is considered as a request for the table of veteran status crossed with sex, so the system evaluates that table's totals to determine whether to suppress Table 3.

The explicit universe definition was too permissive, allowing users to reconstruct a substantial portion of many microdata records. We made the table of sex by age, as in Table 1, for each tract in the U.S for the 2009-2013 ACS 5-year data, and then made the analogue of Tables 3 and 4 for each of 21 target variables. The table of sex by age produced 12,420 uniques nationwide for which Table 1 would not be suppressed. Under the explicit universe definition, anywhere from three to 17 of the 21 variables could be recovered, with a mean of 9.5. Under the implied universe definition, no more than four variables were ever recovered, and usually no variables were recovered. Different pairs of initial variables led to different results, but the explicit universe rules frequently allowed too much reconstruction of microdata. We concluded that it was necessary to use the implied universe rules.

We expect a pair of variables to be particularly vulnerable to a differencing attack if the cells of their cross-tabulation are fairly close to uniformly distributed but there are also a few categories that are rarer than the others. Tables of this sort are likely to pass the acceptance rules but are also good candidates for having uniques. Sex by age is an example of such a table. Our age recode had 14 categories: ages in five-year ranges (0-4, 5-9, etc.) up to 35, except that the 15-19 range was split into a 15-17 range and an 18-19 range; ages in 10-year ranges (35-44, 45-54, etc.) from 35 to 85; and a category for 85 and above. The recoded age variable and the sex variable are relatively close to uniformly distributed in many geographic regions. Thus, the tables necessary to execute the attack usually pass. However, age has a few categories that are less common: the two narrow ones around 18 and the 85 and older group. These categories make uniques more common than if the variable were completely uniformly distributed, and 87% of uniques in tract-level tables that pass fall into these three age categories. The variables most frequently recovered were defined for most or all of the population and had few and relatively evenly distributed categories.

The attack was even more successful when the initial two variables were sex and married, spouse present (MSP). MSP has six categories, of which three—now married, spouse absent; widowed; and separated—occur relatively infrequently. The rarer categories make this variable especially risky, accounting for 99.6% of uniques in tract-level tables that pass.

One way to reduce risk is to combine a variable's riskiest categories with less risky categories. For the attack using sex by age, we can reduce the number of uniques with at least one recoverable variable by 79% by combining 18-to-24-year-olds into a single category and people age 75 and older into a single category. However, this strategy is laborious, requiring the Census Bureau to examine each variable separately to determine the risky categories and to anticipate which variables users would cross-tabulate with each other. Furthermore, if a user planned an analysis where a cutoff at age 20 or 85 was particularly important, combining categories would limit the system's usefulness for that purpose.

The results indicate that the table suppression approach we considered would not adequately protect the data in the MAS.

3.2 How often the MAS provides data under a table suppression approach

The AQS acceptance rules were designed for the decennial census, but the research above shows the rules are too risky for the ACS. A complementary question is how often tables would pass these rules. If tables fail the acceptance rules frequently, then the user will be frustrated and the system cannot be as useful as we might hope.

To test how often a table passed, we made table shells for tables of each of the 23 variables used in Section 3.1, for each two-way combination of 16 of the variables and for each three-way combination of 15 of the variables. By "table shell," we mean the table definition in terms of variables and categories but for no defined geographic region. These tables often contained structural zeros, cells that had to have a count of zero because of the way the variables are defined. For example, the Census Bureau records employment-related variables only for people age 16 and older; other people are out of universe. Thus, a table cell for unemployed people ages 10-14 would by definition have a count of zero. We omitted tables with structural zeros from this analysis, leaving about 500 table shells. The universes for these tables varied; they could include all or only some respondents.

With acceptance rules based on small cell counts, we expect that tables with more cells should pass less often, as the respondents are more spread out across cells. Figure 1 depicts this relationship, showing the probability of passing versus the number of table cells. The points represent the individual table shells, and the curve shows a moving average, using a normal kernel smoother with standard deviation 3. The plot shows that although some table shells have very low pass rates even at the state level, most tables pass, with a noticeable minority of failures when the number of cells approaches 100. Tables with well over 100 cells are possible, but are not shown in Figure 1.

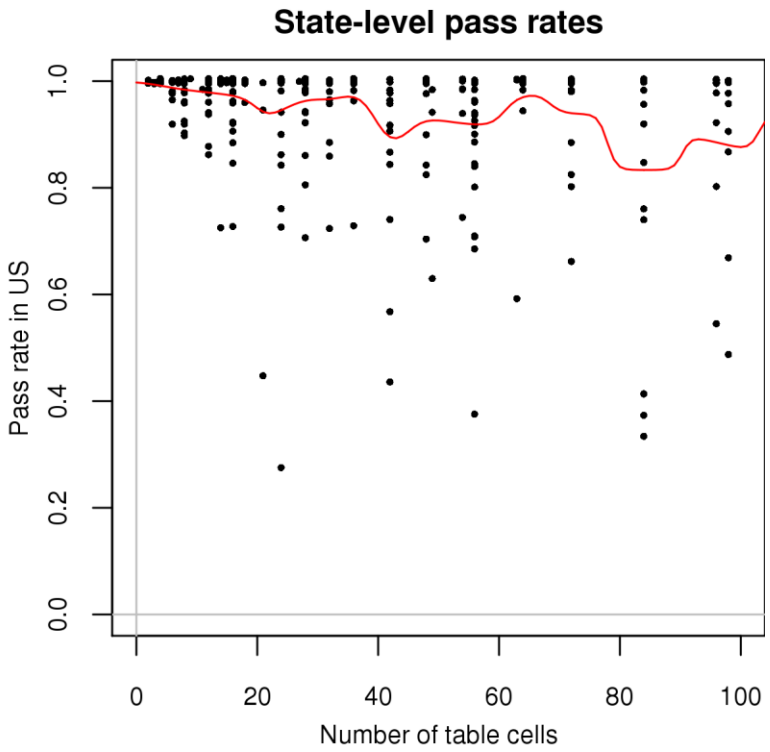

Figure 1: Frequency with which state-level tables pass disclosure rules, plotted against number of table cells.

Figure 2 shows that county-level pass rates are much lower. A very small table will usually pass, but may occasionally fail the acceptance rules, and pass rates drop quickly as the table gets larger. Tables with even 20 cells fail more often than they pass, and tables with 40 or more cells fail dramatically more often than they pass.

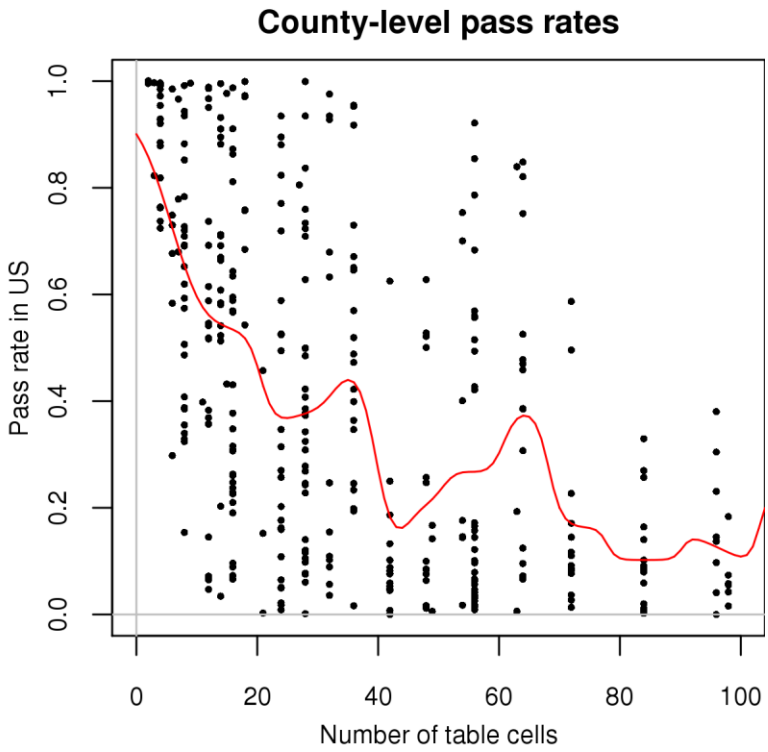

Figure 2: Frequency with which county-level tables pass disclosure rules, plotted against number of table cells.

Figure 3 shows that tract-level tables pass even less frequently. Only about one-third of tables with about 10 cells pass, and tables with more than 40 cells almost never pass. Very small tables usually pass, but fail a significant minority of the time. At the tract level, tables will be suppressed frequently unless they are very basic.

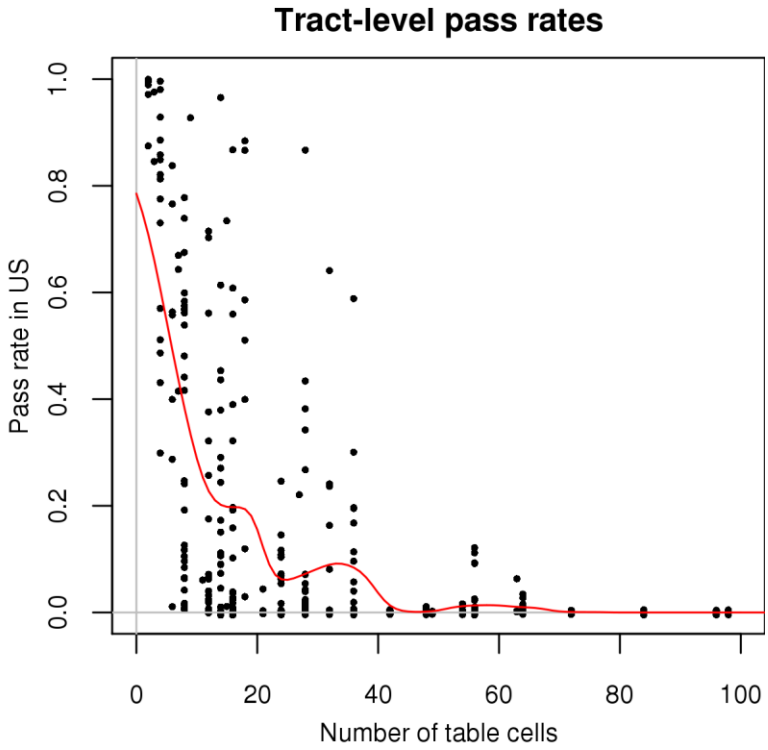

Figure 3: Frequency with which tract-level tables pass disclosure rules, plotted against number of table cells.

Whether a table passes also depends on the dimension of the table, because higher-dimensional tables often have more cells. For each variable combination used in Figures 1 through 3, we constructed the corresponding table for each tract in the country. Figure 4 gives the cumulative density function of pass rates for one-way, two-way and three-way table shells, with the one-way cdf curve being closest to the bottom of the graph and the three-way cdf curve being closest to the top. One-way tables often pass, although some have a very low pass rate. Two-way table shells' pass rates vary, but a majority pass less than 50 percent of the time. Three-way tables pass infrequently at the tract level, with more than three-quarters of table shells having pass rates less than 20 percent. Three-way tables in particular exhibit pass rates low enough to limit the system's usefulness.

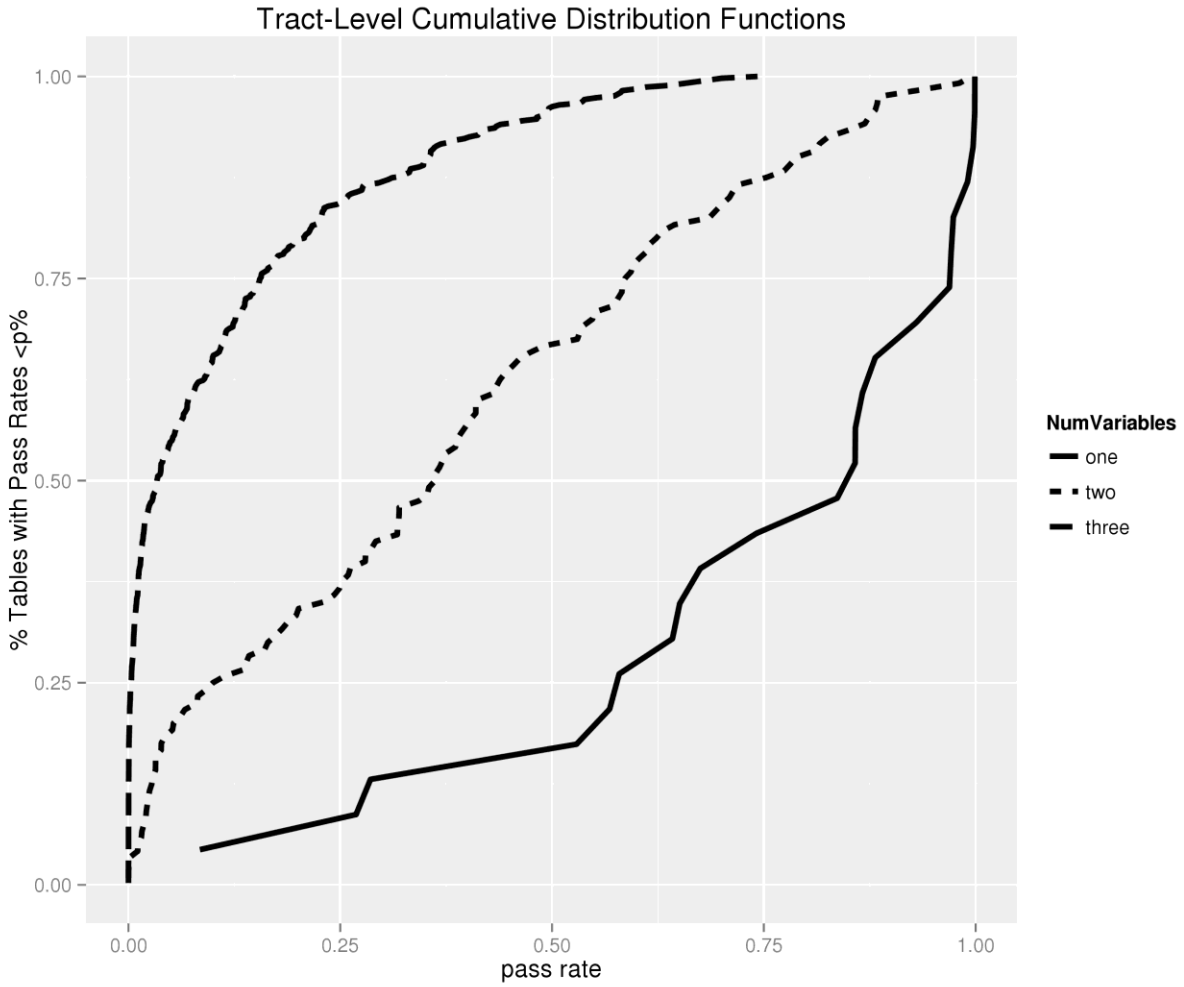

Figure 4: Cumulative distribution functions of table pass rates for one-, two- and three-dimensional tables at the tract level.

Counties are usually much more populous than tracts, so most one-way table shells have a pass rate close to 100 percent, as shown in Figure 5. More than 30 percent of two-way table shells and more than 75 percent of three-way table shells fail the disclosure rules more often than they pass, even at this larger geographic level. Figure 5 further illustrates that, except for simple tables, users would often not get their desired output under these rules.

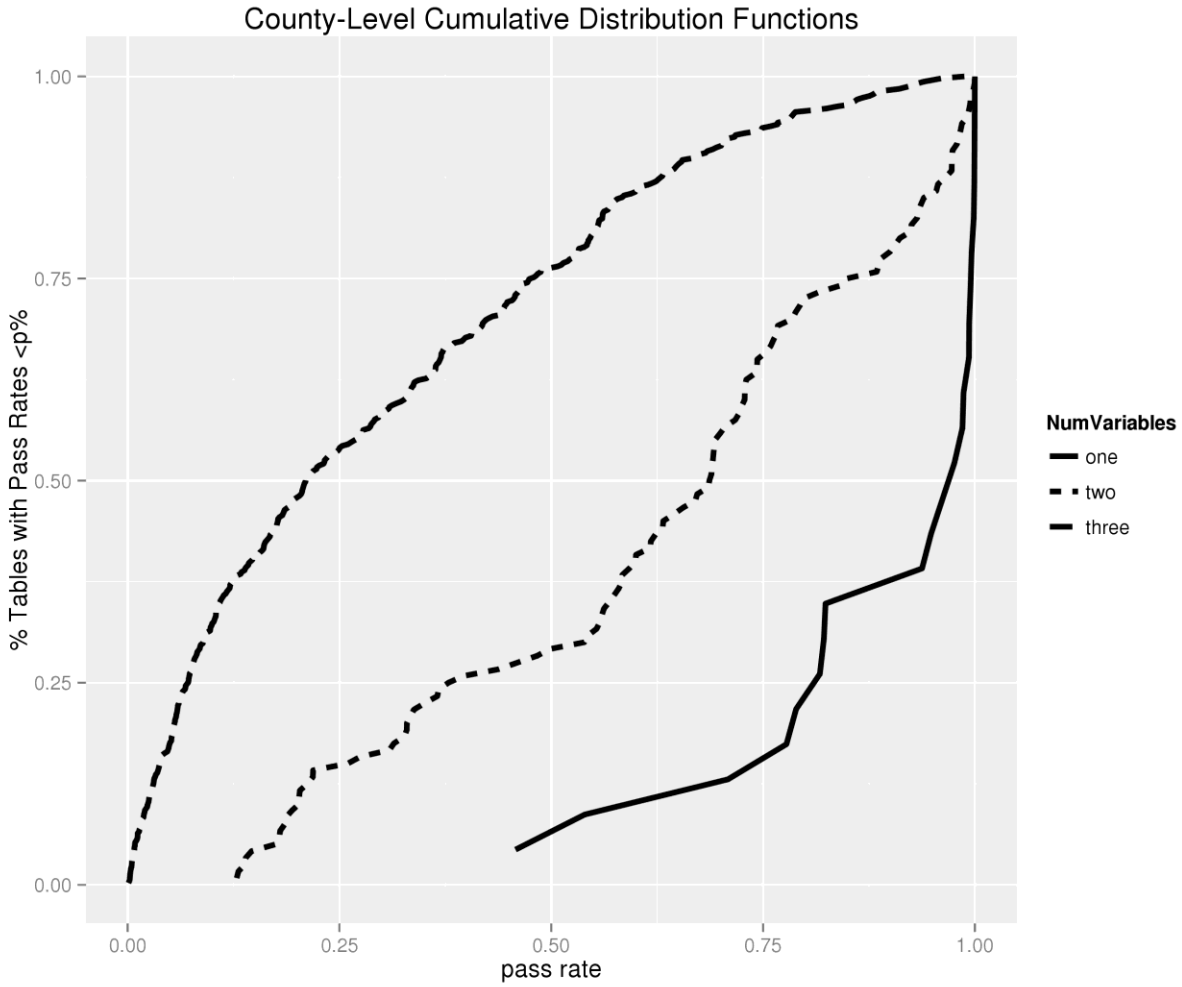

Figure 5: Cumulative distribution functions of table pass rates for one-, two- and three-dimensional tables at the tract level.

3.3 Geographic clustering and composite geographies

Since a table for a composite geography passes only if the table would pass for each component geography, composite geographies pass even less often than Figures 1 through 3 indicate. We expect that users are mostly interested in sets of contiguous tracts, often within the same state. Passing and failing tracts for a given table shell exhibit geographic clustering. Hence the multi-tract areas that users would most often construct pass slightly more frequently than they would if tracts passed or failed independently, but the difference in pass rates is only a few percentage points at best. To test the pass rates, we chose four table shells whose single-tract pass rates varied from very high to very low and constructed the corresponding table for over 70,000 intrastate collections of k contiguous tracts, for k=2,3,4,5. For this small set of table shells, we never observed more than a 4.1 percentage point difference between pass rates for multiple contiguous tracts in the same state and pass rates for the same number of tracts chosen at random from the U.S. The largest difference in this group was for the race

variable with seven categories tabulated over three tracts. As Table 5 shows, more than one-third of the difference in pass rates for collections of three tracts results just from tracts being in the same state, without considering contiguity.

Table 5: Race variable – probability of three tracts all passing

Condition	Probability of passing
Random tracts from US	7.1%
Random tracts from same state	8.7%
Contiguous tracts from same state	11.2%

* The first two estimates are based on all tracts or combinations of tracts in the US. The third estimate is based on a stratified sample of 72,421 intrastate groups of contiguous tracts and has a standard error of approximately 0.1%.

Figures 6 and 7 show which tracts pass and fail the disclosure rules for tables of race in California and Texas, respectively, excluding tracts without population. These two maps show that the passing tracts tend to border each other more often than if the maps were colored in randomly. The clustering seems more visually striking than practically relevant in increasing pass rates for multiple tracts.

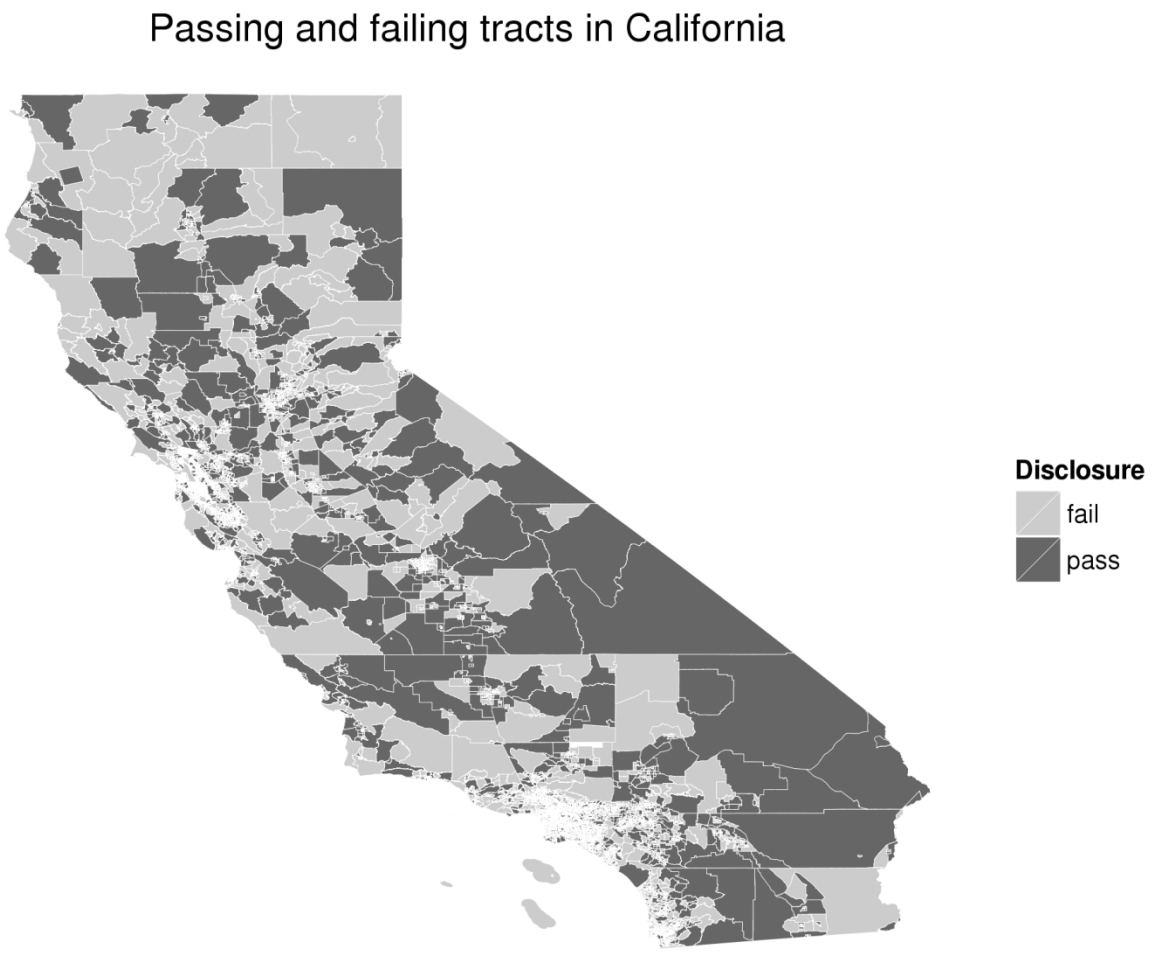

Figure 6: Passing and failing tracts in California for a table of race

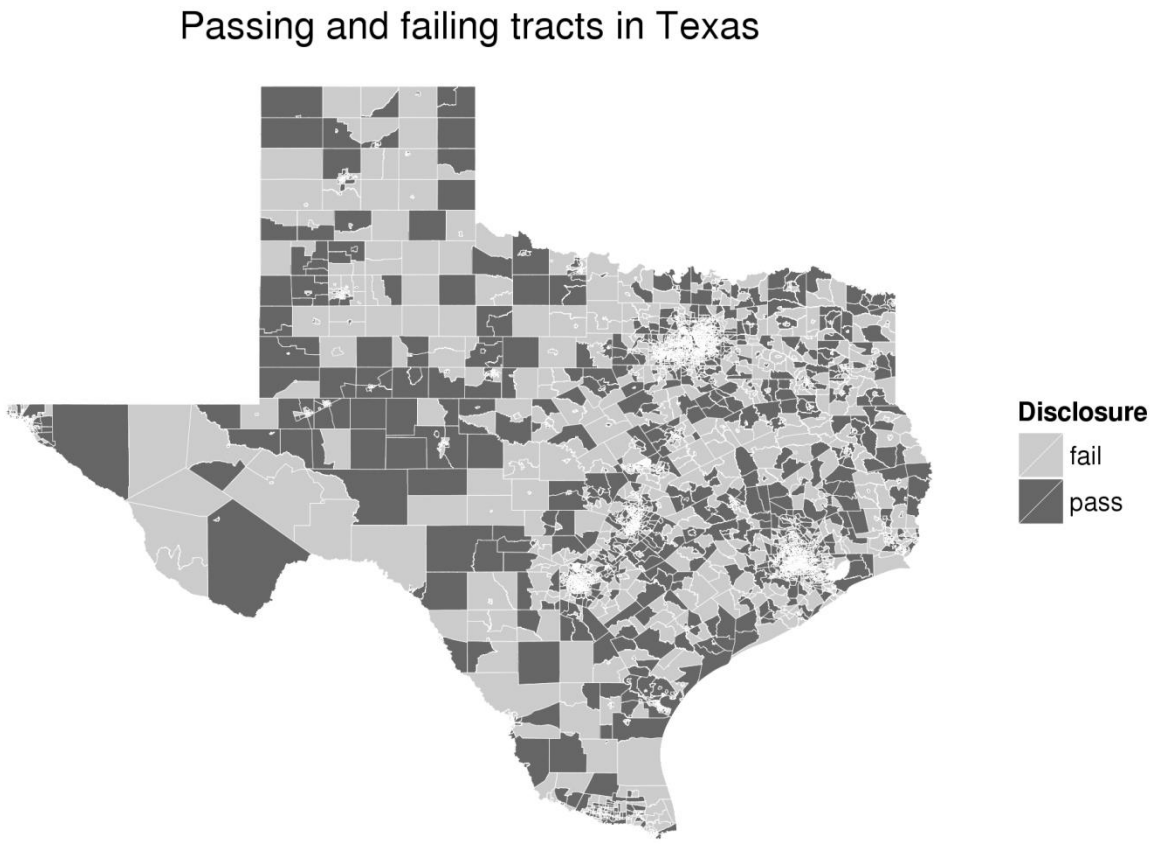

Figure 7: Passing and failing tracts in Texas for a table of race

4. The tradeoff between proportion of tables produced and risk of showing a unique

In any scenario where data are released, the data producer should maximize the data's value to a legitimate user while keeping risk within acceptable bounds. Data with less detail or with modification for disclosure have lower risk than other data, but they are also less useful. For the MAS, the system's usefulness depends not primarily on the quality of the data, but on whether the system releases the data at all.

As we saw with the differencing attack, the most vulnerable units in a dataset are the uniques. Thus, the proportion of uniques that are in non-suppressed tables can be a proxy for disclosure risk. We measure risk relative to the case where the system does not suppress any tables. Not every unique is necessarily risky, nor is every risky unit a unique, so our risk metric is not perfect, but it can give us an idea of how well the disclosure rules work.

Our utility metric is the proportion of tables the system provides. This metric is not the only one possible, as certain tables may be more useful than others to a legitimate user.

Figure 8 shows the relationship between risk and utility if a table of age is made for every tract in the U.S., using the same 14 categories as before. The proportion of uniques that are in released tables is on the x-axis, while the proportion of tables the system releases is on the y-axis. We considered four

acceptance rules with 11 parameters for each rule, ranging from suppressing every table to suppressing no table. The rules were:

- a cap on the percentage of uniques within a given table
- a minimum allowable population
- a minimum allowable median unweighted cell size
- a minimum allowable mean unweighted cell size

The solid line shows the pass rate and proportion of uniques for the percentage of uniques rule as the parameter is varied, and the different styles of dotted lines give the pass rate and proportion of uniques for each of the other rules, taken individually. The gray points in the plot give the same information when the rules are combined in all 11^4 possible ways.

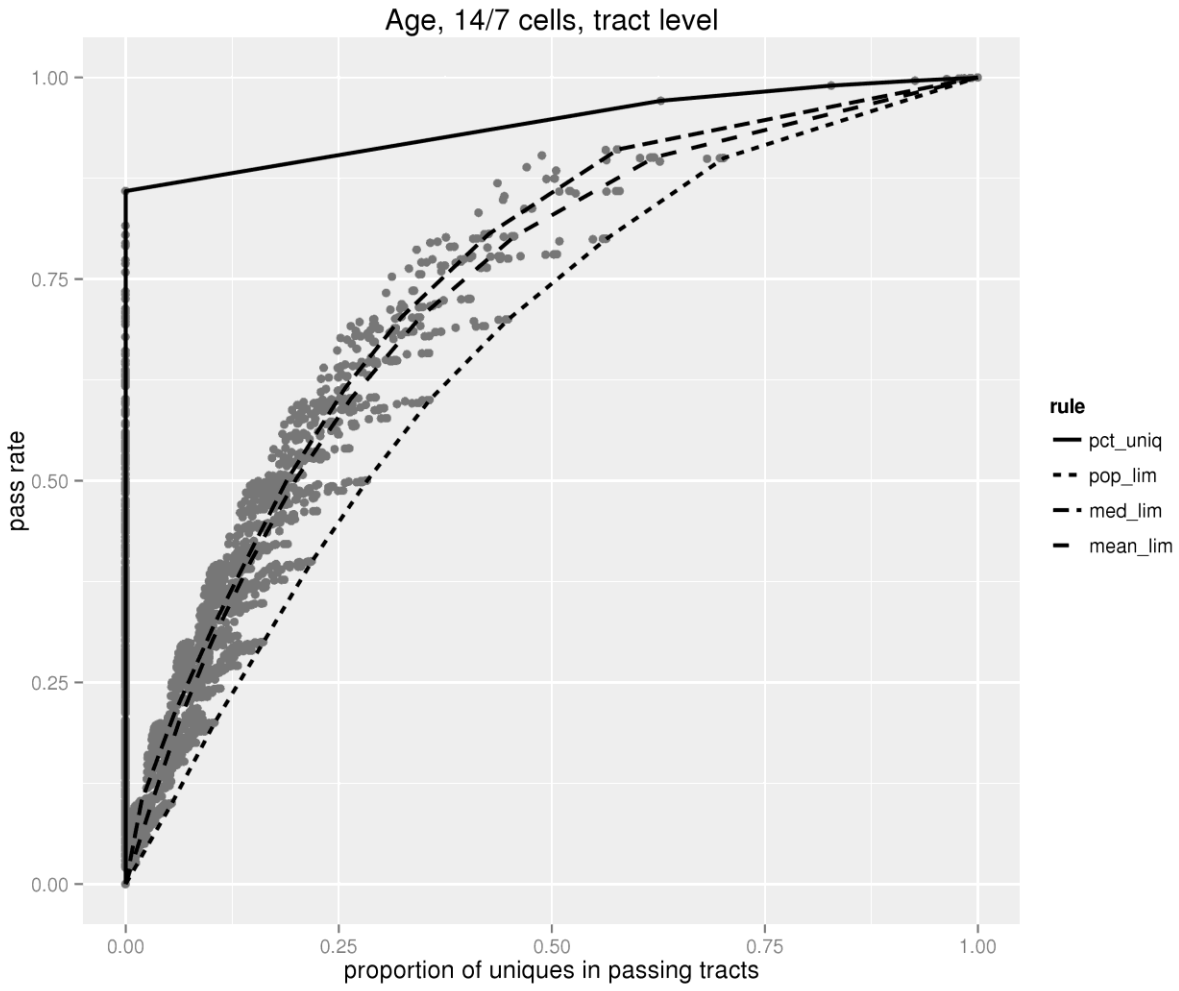

Figure 8: Risk-utility plot for a table of age at the tract level under differing disclosure parameters.

Ideally, the points plotted should lie well above the 45-degree line, indicating that by setting the disclosure parameters to accept minimal risk, we can get a large proportion of the system's potential usefulness. Points above the 45-degree line indicate the graphs the system releases tend to have fewer uniques than the graphs the system does not release.

A less pleasant possibility, which we call the "worst reasonable case," occurs when the points lie on or near the 45-degree line, indicating that the tables released under even strict acceptance rules have roughly as many uniques as those released only under lenient rules. In this case, the rules appear somewhat arbitrary, with the tables that fail exhibiting collectively approximately as much risk as the tables that pass.

An even more problematic case would arise if the points tended to lie below the 45-degree line, creating a convex up shape. This pattern would indicate that stricter rules tend to allow tables with more uniques and suppress tables with fewer uniques. Ideally, this phenomenon would not happen.

The solid curve is also the best possible risk-utility balance under this measurement framework. The curve shows the distribution of uniques: the closer it is to the line $y=1$, the more concentrated the uniques are in relatively few tables, and the more uniques we can suppress by withholding only these few tables. In Figure 8, the solid line is close to the top of the graph, indicating we can eliminate most uniques at a low cost, although the acceptance rules other than proportion of uniques do not maximize this potential.

If uniques are evenly spread over all tables, the solid curve coincides with the 45-degree line. In this case, utility cannot be better than proportional to risk.

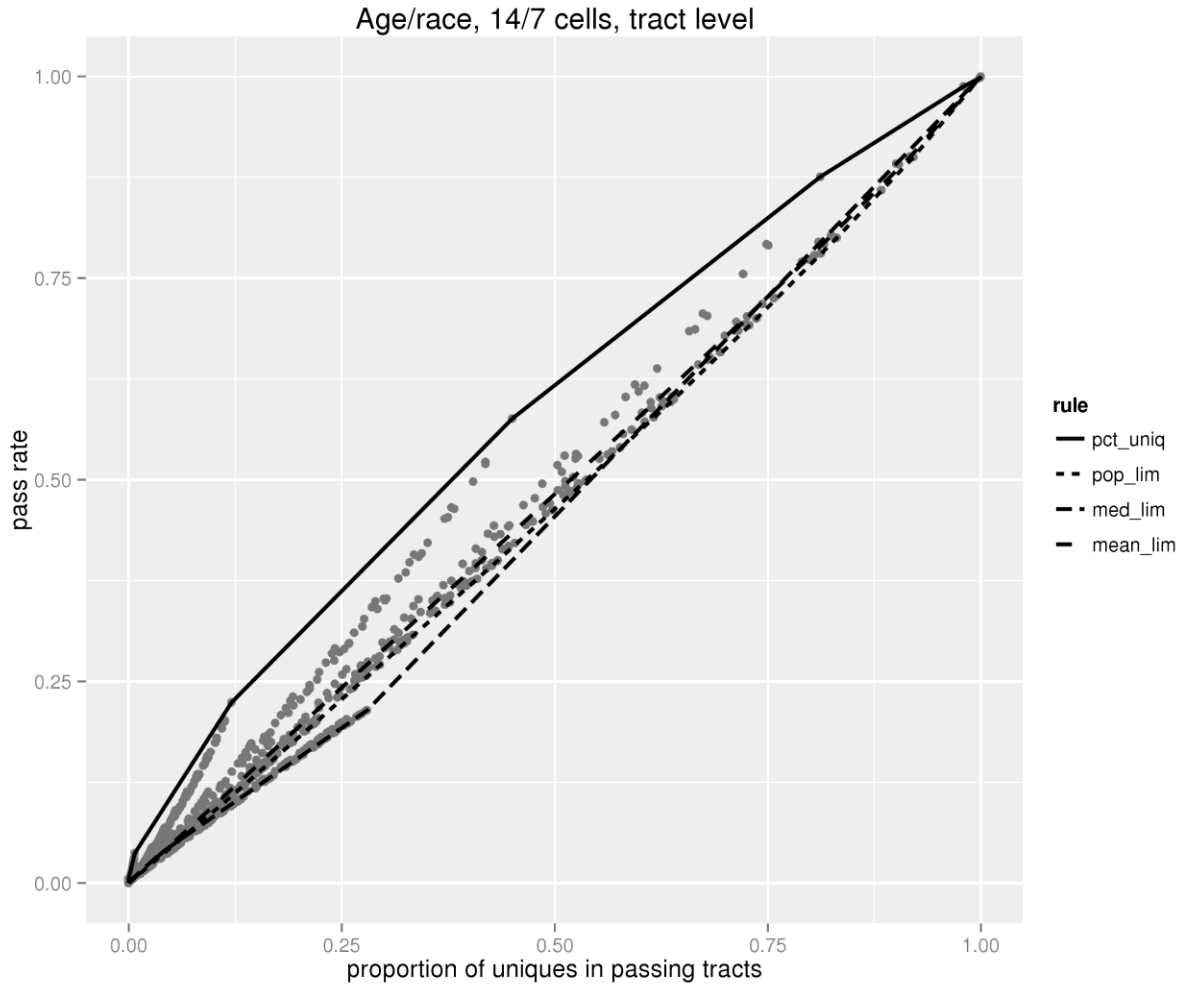

Figure 9: Risk-utility plot for a table of age cross-tabulated with race at the tract level under differing disclosure parameters.

Figure 9 shows a plot similar to Figure 8 when age and race are cross-tabulated at the tract level, for all tracts in the U.S. All three curves other than the solid one fall mostly or entirely below the 45-degree line, worse than the worst reasonable case. The tables released when the acceptance rule is relatively strict contain uniques slightly more frequently than those released only when the acceptance rule is lenient, when any one of the mean, median or population rules is used by itself. The scattered dots indicate that when multiple rules are used in combination, the results are sometimes a little better than the worst reasonable case. No matter what rules we choose, we are hamstrung by the fact that uniques are relatively evenly scattered across all tracts, as shown by the solid line's low position.

The veteran status variable creates an odd pattern, shown in Figure 10. The high solid curve indicates that many tables are obtainable with almost no risk. However, if we use the mean, median or population threshold rules, the risk-utility curves are steepest at the left and right of the graph. This pattern indicates the highest concentration of uniques is in the tables the system is moderately likely to release, more so

than the tables that are either most or least likely to be released. The large gap between the solid curve and the other curves indicates that these rules provide a far worse balance between risk and utility than is theoretically possible. The corresponding graph for employment status taken at the county level has a similar shape.

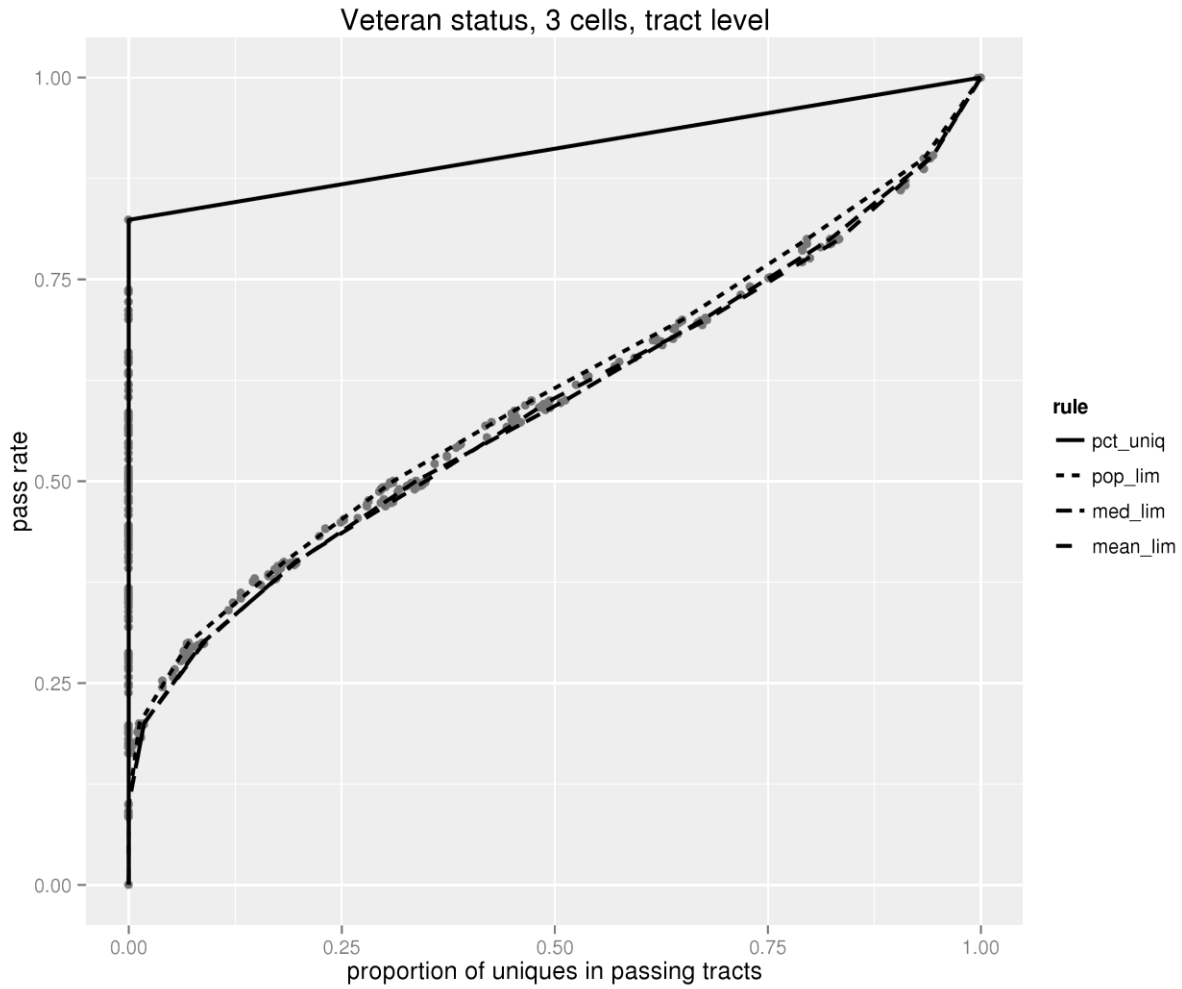

Figure 10: Risk-utility plot for a table of veteran status at the tract level under differing disclosure parameters.

Figure 11 shows an extreme case, the plot for tract-level tables of whether a person lives in a household or group quarters (GQs). The table has few uniques, roughly one per 100 tracts, so even if every table with a unique is suppressed, almost all tables can still be provided. The solid curve adheres closely to the graph's left and upper sides. However, the mean, median and population rules reveal many uniques even while holding back many tables. (The mean and median cell sizes of a two-cell able are equal; hence the curves for the associated rules coincide.) The balance of uniques revealed and tables given is not much better than the worst reasonable case. In this case, the mean, median and population rules do not balance well

the requirements of providing data and preventing disclosure. The corresponding graph for employment status has a similar shape.

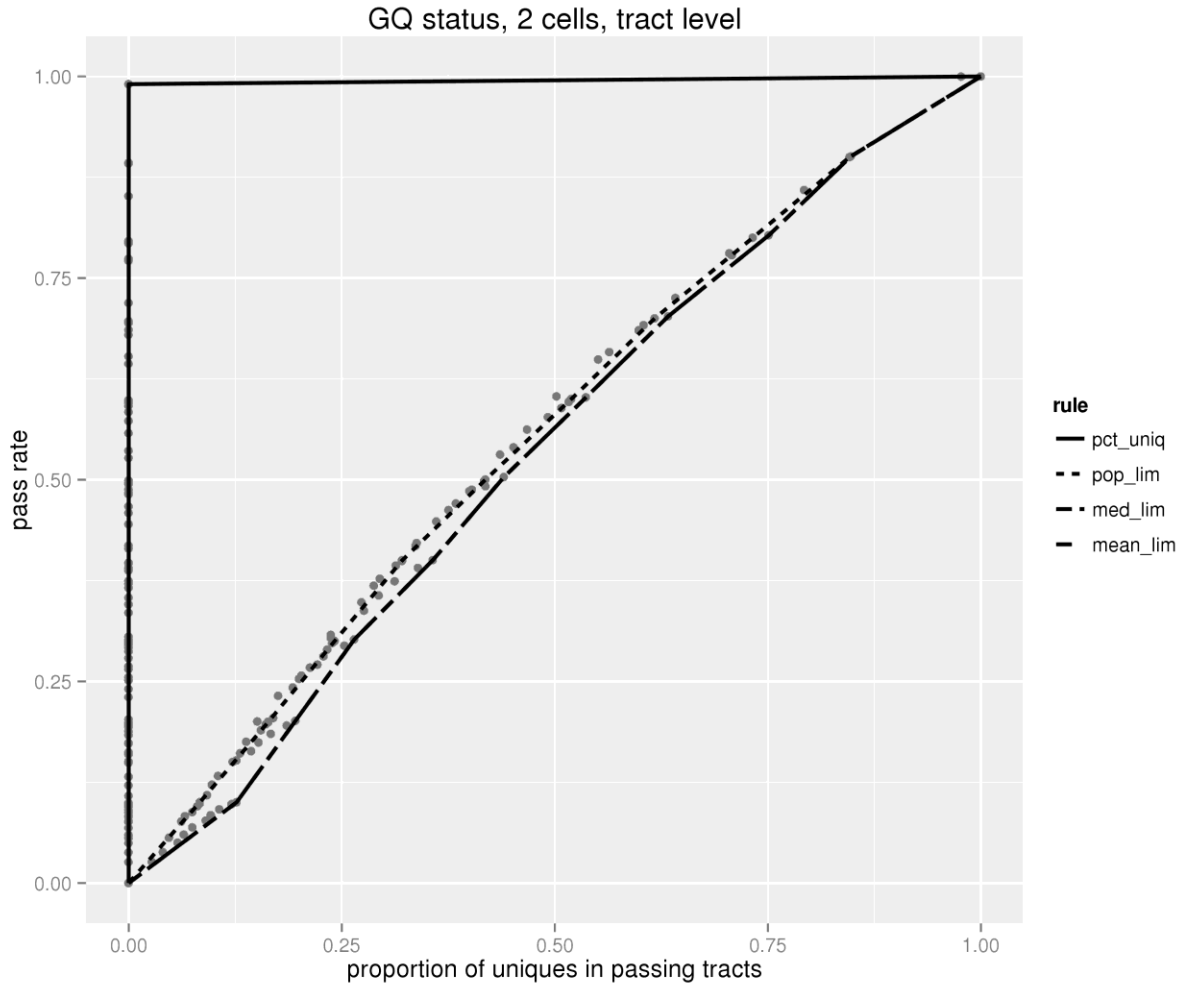

Figure 11: Risk-utility plot for a table of group quarters status at the tract level under differing disclosure parameters.

From these graphs, we conclude that variables or combinations of variables behave substantially differently in terms of how the number of uniques (loosely speaking, risk) increases as more tables are released. For some table shells, many tables can be released with little or no risk. For other table shells, the risk is roughly proportional to the number of tables released. In addition, the mean, median and population rules, taken individually and in combination, generally do not come close to giving the optimal balance between tables released and uniques shown, and occasionally these rules give a worse balance than simply selecting tables at random to release. If we pursued the MAS project further, we could consider tailoring the rules depending on the behavior of a given variable, but this approach might require us to examine the behavior of every variable combination individually and lose the benefits of automation.

5. Alternative approaches

Our research shows the disclosure rules for the AQS do not adequately protect the ACS data and may lead to the disclosure of an individual's attributes or identity. These rules would also cause the MAS to deny requests for many tables. Tightening the rules would lessen the disclosure problems but would result in even fewer released tables. We would end up with a system where a user only rarely receives the requested table.

We concluded that a remote access system producing tables based on the internal microdata file cannot be sufficiently useful and sufficiently safe for the ACS. Thus, we have stopped working on the MAS and considered other ways to make our data available. Although the Census Bureau is considering a number of options to make data accessible while preserving confidentiality, our focus has shifted from queries with post-tabulation suppression to disclosure methods directly protecting the source data, with the hope that the protected source data may even be releasable directly to the public. Synthetic data appears to be the most promising method and may be used for purposes that go beyond our original intentions for the MAS.

5.1 Synthetic data

Synthetic data is the approach of using a model to generate data designed to be similar to the collected data (Rubin, 1993). Synthesis may be full, meaning that the whole dataset is generated, or partial, meaning that some records or variables are generated and the original data are used for the rest of the dataset (Reiter, 2003). In a fully synthetic dataset, there need not be any direct correspondence between a simulated record and a record in the original sample.

Synthetic data could be used for a variety of data products. First, they could be incorporated into the official microdata used to create all derivative data products. For instance, the ACS already uses synthetic data to protect group quarters data. Second, they could be used in more cases of public-use microdata, as they are for the Survey of Income and Program Participation Synthetic Beta (SSB) and the Synthetic Longitudinal Business Database (SynLBD) (U.S. Census Bureau, 2016; U.S. Census Bureau, 2013b). Third, based on future research, they could be used as source data in an online system. OnTheMap is such a system already. Future research could determine if a newly envisioned system, with synthetic or perturbed data as a source, could be created for demographic surveys.

Because a model cannot preserve every possible feature of a dataset, particularly a large one such as the SSB or the SynLBD, the Census Bureau offers validation of any results obtained by outside researchers from these two files. Researchers can provide code to the Census Bureau to check results against the unsynthesized confidential data used to create the SSB and the SynLBD. The SSB website warns that "Without validation of results, Census, SSA, and IRS make no guarantee of the validity of the SSB for any research purpose" (U.S. Census Bureau, 2016). The SynLBD website states that "Validating all possible relationships between SynLBD variables has not been feasible" (U.S. Census Bureau, 2013a). The validation service and the accompanying disclaimers illustrate the challenges of creating synthetic data for large datasets with many variables.

The Census Bureau also produces the LEHD Origin-Destination Employment Statistics (LODES) data, which show where people live and work. These data are partially synthetic and may be downloaded (U.S.

Census Bureau, 2015b) or accessed through the Census Bureau's OnTheMap application (U.S. Census Bureau, 2015c).

6. Conclusion and next steps

A table server protected primarily by table suppression is not a viable option for the ACS. Synthetic data seems to be the most promising approach, and we have begun the further research necessary to evaluate its effectiveness, both in providing meaningful data and in protecting respondent confidentiality.

The Census Bureau is beginning work on products that will accomplish those goals of the MAS that do not cause disclosure risks. We plan to produce a table aggregator that will provide exact variance measures for composites of geographies or variable categories, which have so far been unavailable. If we produce a set of synthetic microdata, it may be accompanied by a table compiler as a convenience for users who want to create tabulations without manipulating the data themselves.

Our research on the MAS serves as a reminder of the challenges that arise when data owners attempt to release information with minimal pre-tabulation disclosure protection. During the research, we developed new methodology for testing the usefulness and vulnerabilities of a query system. We learned about what is possible and what is not in safely releasing data, and we will build on this knowledge in developing other disclosure methods.

References

Australian Bureau of Statistics. (2014a) Remote Access Data Laboratory (RADL). Australian Bureau of Statistics, Canberra, Australia (Available from http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL.)

Australian Bureau of Statistics. (2014b) What is DataAnalyser? Australian Bureau of Statistics, Canberra, Australia (Available from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1406.0.55.006~User%20Guide~Main%20Features~What%20is%20DataAnalyser~1.)

Australian Bureau of Statistics (2015) TableBuilder. Australian Bureau of Statistics, Canberra, Australia (Available from http://www.abs.gov.au/websitedbs/censushome.nsf/home/tablebuilder.)

Blash, R. (2015) Center for Enterprise Dissemination Services and Consumer Innovation. Census Scientific Advisory Committee meeting, April 16, 2015. (Available from http://www2.census.gov/cac/sac/meetings/2015-04/2015-Blash_presentation.pdf.)

Lauger, A., Wisniewski, B. and McKenna, L. (2015) Disclosure avoidance techniques at the U.S. Census Bureau: Current Practices and Research. In *Proc. 2015 Joint Statistics Meetings, Government Statistics Section*, pp. 3630-3642. (Available from http://www.eventscribe.com/2015/ASA-JSM/assets/pdf/234197.pdf.)

LIS Cross-National Data Center in Luxembourg. (2011) LISSY. LIS Cross-National Data Center in Luxembourg, Luxembourg (Available from http://www.lisdatacenter.org/data-access/lissy/.)

National Center for Education Statistics. (n.d.) DAS Website. National Center for Education Statistics, Washington, D.C. (Available from http://nces.ed.gov/das/.)

National Center for Education Statistics. (n.d.) NCES Datalab. National Center for Education Statistics, Washington, D.C. (Available from http://nces.ed.gov/datalab/.)

Meyer, P. S. (2014) "Virtual data access" for statistical and research purposes. FCSM Statistical Policy Seminar, December 15, 2014. (Available from http://www.copafs.org/UserFiles/file/2014fcsm/06_PeterMeyerFCSM%20120314.pdf.)

Meyer, P. S. (2015) NCHS remote data access systems: present and planned. FCSM Research Conference, December 3, 2015.

Reiter, J. P. (2003) Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-188.

Reiter, J.P. (2005) Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21, 441-462.

Simard, M. (2011) Real Time Remote Access at Statistics Canada: development, challenges, and issues. In *Proc. 58th World Statistics Congress of the International Statistical Institute*, pp. 3134-3143. (Available from http://www.2011.isiproceedings.org/papers/650096.pdf.)

U.S. Census Bureau (2013a). Synthetic Longitudinal Business Database (SynLBD): analytic validity. U.S. Census Bureau, Washington, D.C. (Available from http://www.census.gov/ces/dataproducts/synlbd/analyticvalidity.html.)

U.S. Census Bureau (2013b) Synthetic Longitudinal Business Database (SynLBD): SynLBD beta data. U.S. Census Bureau, Washington D.C. (Available from http://www.census.gov/ces/dataproducts/synlbd/.)

U.S. Census Bureau (2015a) American Community Survey (ACS) custom tables. U.S. Census Bureau, Washington D.C. (Available from https://www.census.gov/programs-surveys/acs/data/custom-tables.html.)

U.S. Census Bureau (2015b) LODES data. Longitudinal-Employer Household Dynamics program. U.S. Census Bureau, Washington, D.C. (Available from http://lehd.ces.census.gov/data/lodes/.)

U.S. Census Bureau (2015c) OnTheMap application. Longitudinal-Employer Household Dynamics Program. U.S. Census Bureau, Washington, D.C. (Available from http://onthemap.ces.census.gov/.)

U.S. Census Bureau (2016) Synthetic SIPP data. U.S. Census Bureau, Washington, D.C. (Available from https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html.)

Vsevolozhskaya, Olga A., and Anthony, J. C. (2014) Confidence interval estimation in R-DAS. *Drug and Alcohol Dependence*, 143, 95-104.